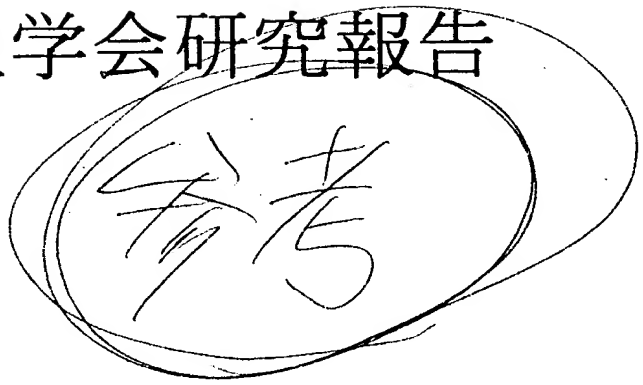
学術刊行物 情処研報 Vol.99, No.20
ISSN 0919-6072

情報処理学会研究報告



99 － FI － 53

1999 年3月1日

社団法人 情報処理学会

情 報 学 基 礎 53－2
(1999. 3. 1)

文書タイプ分類による問題解決向きWWW検索システムの開発と評価

松田 勝志，福島 俊一
NEC ヒューマンメディア研究所

さまざまな問題解決に利用できる文書タイプという概念を導入したWWW検索システムについて述べる．文書タイプとは，一般的な概念構造であるカテゴリーとは違い，特定の問題解決に利用できるコンテンツの種類である．あらかじめこのような文書タイプにWWWページを分類しておくことによって，問題解決に必要なコンテンツを的確に検索することができる．文書タイプへの分類には，WWWページの単語だけではなく，構造的な特徴を利用する．

本稿では，構造的特徴量による分類システムを含む文書タイプ検索システムの詳細について述べ，有用性と実用性を示す実験結果について報告する．大規模なWWWページデータを対象とした実験では，キーワードのみの検索より分類の適合率で40ポイント以上の精度向上が確認できた．

Development and Evaluation of WWW Retrieval System for Problem Solving by Document Type Classification

Katsushi Matsuda and Toshikazu Fukushima
Human Media Research Laboratories, NEC

This paper proposes a novel approach to accurately searching web pages for relevant information in problem solving by specifying a web document category instead of the user's task. Accessing information from World Wide Web pages as an approach to problem solving has become commonplace.

To specify a user's problem solving task, we introduce the concept of document types that directly associate with the problem solving tasks and are easily designated by users in our system. The system classifies web pages into the document types by comparing their pages with typical structural characteristics of the types. In this paper, we report details of our developed system and experimental results. The average precision of the system is 80% or more. In comparison, the average precision without classification is less than 30%.

1. はじめに

WWW(World Wide Web)の普及に伴い，問題解決向けの様々な検索サービスがインプリメントされてきている．これらのサービスは従来の汎用のキーワード検索ではなく，特定の分野やタスクに特化した精度の良い検索サービスが多い[1][2][3]．一般ユーザにとって的確なキーワードを決めることは困難なため，汎用のキーワード検索では検索結果にゴミが多くなってしまう．今後はAskJeeves[4]のような特定の分野やタスクに特化した検索サービスが増えてくるであろう．

筆者らが開発した問題解決向けWWW検索システムでは，文書タイプという概念を導入した．文書タイプとは，ディレクトリサービスで使われる一般的な概念構造であるカテゴリーとは違い，特定の問題解決に利用できるコンテンツの種類である．このシステムでは，WWWページをあらかじめその文書タイプに固有の構造的な特徴をもとにそれらの文書タイプに分類する．そして検索時にユーザが問題解決の種類に応じた文書タイプを指定することで的確な検索ができる．

本稿では，まず筆者らが導入した文書タイプという概念とシステムについて例を示して述べ，また実験結果を示すことによって本システムの有効性と実用性を明らかにする．

2. 検索サービスの問題点

World Wide Webにおいて求めるページを的確に検索するためには，検索キーワードを増やす方法が一般的である．しかし，goo[5]等に代表される従来のキーワード検索サービスでは，平均2個弱のキーワードしか使われないという調査報告[6]があり，一般のユーザにとってキーワードを増やすことは難しい．

一方，人手によって分類されたカテゴリーを辿ることによって検索機能を提供するディレクトリ検索サービスがある．代表的なサービスにYahoo Japan[7]がある．カテゴリー木構造を辿ることで検索することができるため，精度の良い検索が行えるが，検索対象ページが少ないであるとか，更新に時間がかかりすぎる等の問題点もある．これに対処するために計算機によってカテゴリーに分類する研究がなされている[8][9][10]．しかし，カテゴリー構造は一般的な概念体系に則り構築されているため，求めるコンテンツが含まれたページを容易に見つけることができるとは言い難い．

検索サービスを利用するユーザは，何らかの問題解決を行うことを目的としていることが多い．例えば，パソコン購入や旅行計画などである．実際これらの問題解決に役立つオンラインショッピングやトラベルのサイトへのリンクを集めたポータルサイトもある[11]．しかし，このような問題解決すべてについて質の高いリンクを用意するのはコストの面から非常に困難である．また，問題解決を行うことを目的としているユーザは，その問題解決に即したキーワードを想定することはできても，一般的な概念体系に則ったカテゴリー構造中にそのキーワードを見つけ出すことは困難である．そもそもそのようなキーワードで分類されていないことも多い．例えば，パソコンの購入計画や改造では『カタログ』や『オンラインショップ』，就職状況の調査では『求人案内』，英語論文作成では『国際会議』などである．これらのキーワードはカテゴリー構造にはなく，カテゴリー構造を横断していることが多い．

このように問題解決にWWWのコンテンツを利用したいユーザの検索を考えた場合，従来の検索サービスでは的確な検索ができないであるとか構築のコストがかかる等の問題が表面化してくる．すなわち，低コストで構築でき，問題解決に利用できるページをユーザが的確に検索可能な検索サービスが必要である．

3. 文書タイプ分類と検索

本節では，従来の検索サービスでは困難な問題解決のためのWWW検索が容易に可能な検索システムのアイデアとそのシステム構成について述べる．

3.1 文書タイプ

前節で述べた問題解決タスクに即したキーワードである『カタログ』，『オンラインショップ』，『求人案内』や『国際会議』等について考えてみる．これらのキーワードはある問題解決から比較的容易に想定可能である．上記のキーワード以外にも，例えば，料理という問題解決ならば『料理レシピ』などが想定できる．また，ユーザはあるWWWのページを見た時，そのページが『商品カタログ』であるとか『求人案内』であることを一瞥しただけで判別することが可能である．そこで筆者らは，ある問題解決にはその問題解決に応じて要求されるコンテンツのタイプがあり，そのタイプはある種の固有なページのスタイルを持っているのではないかという仮説を立てた．例えば，購入計画という問題解決にはカタログのようなものが要求され，カタログにはそのページがカタログであると同定できる固有なスタイルを持っている，ということである．このタイプを文書タイプと呼ぶ．

文書タイプは，ユーザがある問題解決に直面した場合，容易に指定可能でなければならないため，慎重に選ぶ必要がある．以下に，ビジネスユース，パーソナルユースでの文書タイプの例を示す．

ビジネスユース	パーソナルユース
カタログ	
オンラインショップ	
FAQ	
リンク集	
調査報告	料理レシピ
求人案内	プレゼント
事例	教室・講座
イベント情報	アップデートプログラム

表1．文書タイプの例

これらのような文書タイプをあらかじめ用意しておくことによって幅広いユーザの検索要求に対応することができる．

文書タイプを指定して検索が可能なように本システムでは，あらかじめ文書タイプにWWWページを分類しておく．

3.2 構造的特徴量による分類

WWWページを文書タイプに分類するには，その文書タイプに固有の構造的特徴を利用する．従来さまざまな文書分類の研究がなされているが，それらは文書中の単語のみに着目したものが多い．しかし，WWWのHTML文書にはさまざまな付加情報(タグ，イメージ，ハイパーリンク等)が内包されている．実際にユーザはWWWのページを一瞥するだけでそのページが『カタログ』であるか『求人案内』であるかということが判断できる．これはそれらの文書タイプに応じたデファクト的なページの形式や最低限の項目や要件等が存在するためである．例えば，『カタログ』であれば，商品名が目に付き易い大きさで表現され，その商品の画像があり，仕様や特長を記した部分またはページへのリンクがある，等である．人はこれらの情報をもとに総合的に判断する．

本システムの分類では，人が判断するのと同様にWWWページをある文書タイプに典型的なタグと文字列，インラインイメージのサイトや数，リンクの種類と数，URL自身の文字列などの構造的な特徴から総合的に判断して分類する．

図1に分類システムの構成を示す．

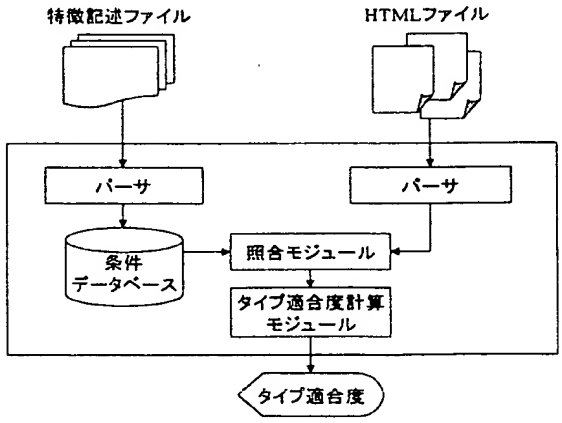


図1　分類システムの構成

分類には，文書タイプ毎に用意したその文書タイプの典型的な構造的特徴を記述した特徴記述ファイルを用いる．特徴記述と収集したWWWページ群とをパターンマッチし，各ページのその文書タイ

プへのタイプ適合度を求めることで分類を行う．

文書タイプ1種類につき1ファイルの特徴記述を用意する．特徴記述ファイルに記述される条件には，現在以下の5種類がある．

(1) keyword
(2) image
(3) link
(4) url
(5) structure

(1)のkeyword条件は，タグと文字列のペアによる条件である．例えば，『カタログ』文書タイプのページには，「仕様」や「特徴」などの文字列がアンカーとしてある場合がある．keyword条件はこのような特徴とマッチする．keyword条件には以下のようなものがある．

keyword:3:<a>:仕様 / 特徴

この条件式は，<a>タグ中に「仕様」または「特徴」という文字列があった場合，3点を追加する，ということを示している．この3点というのは，この条件の重みであり，特徴記述を作成する人が自由に設定することができる．

(2)のimage条件は，インラインイメージに関する条件である．例えば，『カタログ』文書タイプのページには，その商品の画像イメージが含まれている場合がある．image条件はこのような特徴とマッチする．以下にimage条件の例を示す．

image:2:over(4000)>=1

この条件式は，4,000バイト以上のイメージが1個以上あった場合，2点を追加する，ということを示している．4,000バイトという数値はマジックナンバーであるが，経験的に得られた商品画像イメージのサイズの下限である．

(3)のlink条件は，リンクに関する条件である．例えば，『カタログ』文書タイプのページには，その商品の仕様や特徴やQ&Aに関するページへのリンク(そのサイト内部へのリンク)がいくつかある場合がある．また，『リンク集』文書タイプのページには，そのページの作者が集めた有益なページへのリンク(ほとんどがそのサイト外部へのリンク)が多数ある場合がある．link条件はこのような特徴とマッチする．link条件には以下のようなものがある．

link:1:internal>=3

link:5:external>=15

上の条件式は，『カタログ』向けの条件であり，サイト内部へのリンクが3個以上あった場合，1点を追加する，ということを示しており，下の条件式は，『リンク集』向けの条件であり，サイト外部へのリンクが15個以上あった場合，5点を追加する，ということを示している．

(4)のuri条件は，ページそのもののURL文字列に関する条件である．例えば，『カタログ』文書タイプのページのURLには，その組織種類を示す"co"と商品ディレクトリを示す"product"という文字列が含まれる場合がある．url条件はこのような特徴とマッチする．次にurl条件の例を示す．

url:4:organization=co&path=product

この条件式は，組織種類を示すURLに"co"があり，かつURLのパスに"product"がある場合，4点を追加する，ということを示している．

(5)のstructure条件は，タグ構造そのものに関する条件である．例えば，『カタログ』文書タイプのページには，その商品の仕様を表形式でまとめている場合がある．structure条件はこのような特徴とマッチする．以下にstructure条件の例を示す．

structure:2:<table border=%1>:
3>=%1>=1

この条件式は，境界線の幅が1以上3以下の表がある場合，2点を追加する，ということを示している．

図2に『カタログ』文書タイプを判別する特徴記述の例を示す．

```
keyword:3:<h2>:製品|商品|サービス|システム
keyword:2:<body>:お客さま|お客様
keyword:1:<body>:登録商標|株式会社|
            "All Right Reserved"|Copyright
keyword:3:<a>:仕様|特徴
image:2:over(4000)>=1
link:1:internal>=3
url:4:organization=co&path=product
structure:2:<table border=%1>:3>=%1>=1
```

図2　特徴記述

この特徴記述では，上で例として示した条件の他に『カタログ』と判断するのに有益ないくつかのkeyword条件を加えている．

各文書タイプについて図2のような特徴記述を用意し，WWWページすべてについて各文書タイプへのタイプ適合度を計算する．タイプ適合度は，特徴記述に含まれるすべての条件のうちマッチした条件の点数の合計をすべての条件の点数の合計で割った百分率で表している．例えば，図2の例で，あるWWWページがstructure条件のみマッチしなかった場合は，『カタログ』のタイプ適合度は89%となる．同様にしてその他の文書タイプのタイプ適合度を計算する．その結果，あるWWWページは，『カタログ』のタイプ適合度89%，『オンラインショップ』のタイプ適合度32%，『求人案内』のタイプ適合度0%，『国際会議』のタイプ適合度5%，のように各文書タイプへのタイプ適合度が算出される．

3.3 文書タイプ指定による検索

文書タイプ指定検索システムを構築した．図3にその構成図を示す．

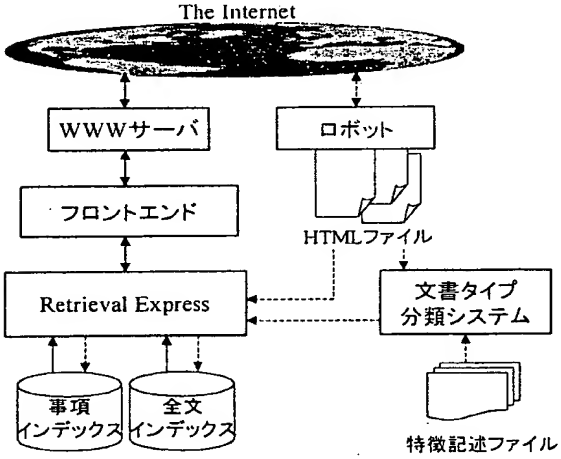


図3　文書タイプ検索システムの構成図

図3において，点線は登録時の処理であり，実線は検索時の処理である．検索システムの単語検索には文字ベース検索エンジンRetrievalExpress(REx)[12]を用いている．登録時，ロボットによって収集されたHTMLファイルは，分類システムとRExにそれぞれ渡す．分類システムの出力結果である各文書タイプのタイプ適合度はRExの事項インデックスに格納する．RExに直接渡ったHTMLファイルはRExの全文インデックス(文字ベースのインバーテッドファイル)に格納する．検索結果は，RExのキーワード検索結果と文書タイプとの論理積をとる．そのため，2つのインデックスでは各HTMLファイルに対して共通のIDを付ける．検索時，フロントエンドは，ユーザの指定したキーワードと文書タイプそれぞれについて全文インデックスと事項インデックスに問合せする．そして，それぞれの結果の論理積を検索結果として表示する．

図4に本システムが搭載された検索システムの画面イメージを示す．図4の検索システムでは，文書タイプとして『カタログ』と『リンク集』が使える．キーワードを入力し，文書タイプを選択することで文書タイプ検索を行う．本システムでは，タイプ適合度が50以上のものを表示するようになっているが，このタイプ適合度の条件は自由に設定可能である[1]．

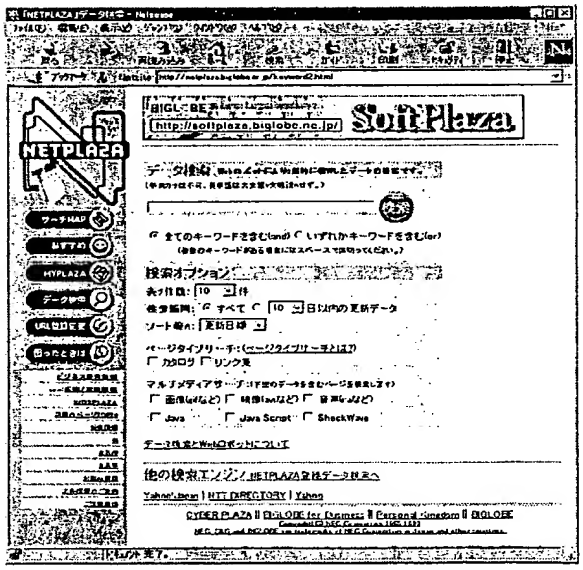


図4　文書タイプ検索システム

4. 評価実験

本節では，本システムが，低コストで構築でき，問題解決に利用できるページをユーザが的確に検索することが可能か

[1] ユーザが自由にタイプ適合度の条件を変更できるようにすることは容易である．

どうかを評価するためにいくつかの実験を行った．

4.1 特徴記述作成と分類速度

本システムで文書タイプを追加する際のコストを調査するために，4種類の文書タイプ(『調査報告』，『求人案内』，『プレゼント』，『アップデートプログラム』)を追加するに費やした時間を調べた．その結果，それぞれの文書タイプを追加するために費やした時間は，特徴記述の作成およびチューニングを含めてのべ5時間程度であった．また，作成した特徴記述ファイルは400～1,000バイト程度であった．

次に，日々増加するWWWページを実時間で登録することが可能かどうかを調べるため，分類速度を調査した．1,000ページ約7.5MバイトのHTMLファイルを文書タイプ6種類に分類するのに費やした時間は，EWS4800/460(CPUはR10000，周波数は200MHz)で約140秒であった．

このように少ないコストで文書タイプを追加することが可能であり，また，十分実用的な速度で分類することが可能であることがわかった．

4.2 分類精度

6種類の文書タイプで分類精度の評価実験を行った．本来なら，ユーザが要求している問題解決に直接利用できるWWWページが検索できるかどうかについて評価する必要がある．しかし，キーワードと文書タイプのみの入力では実際にどのようなコンテンツを要求しているのか判断できない．例えば，パソコンの機種選定という問題解決と価格調査という問題解決を考えてみる．非常に弱い検索条件では，双方の問題解決ともキーワードに"WindowNT"，文書タイプに『カタログ』を指定するであろう．しかし，求めるコンテンツは機種選定の場合は仕様であり，価格調査の場合は価格である．また，同じ機種選定という問題解決であっても，最新機種のみを知りたい場合や旧モデルでも構わない場合がある．このように指定されたキーワードと文書タイプのみではユーザが具体的にどのようなコンテンツを期待しているかが判断できない．また，このような評価のためのテストコレクションもまだ存在しない．そのため，今回の実験では，ユーザが指定した文書タイプへの分類精度という観点で実験した．

4.2.1 実験1

実験1では収集した26,000ページ強のHTMLファイルを対象とした．これに対して全文検索を行い，結果セットを人手によって文書タイプに分類し，それを正解セットとした．『カタログ』，『リンク集』ともに3回ずつ別のキーワードで実験を行い，分類の適合率と再現率を求めた．図5と図6にそれぞれの文書タイプの適合率と再現率の平均値をプロットした．

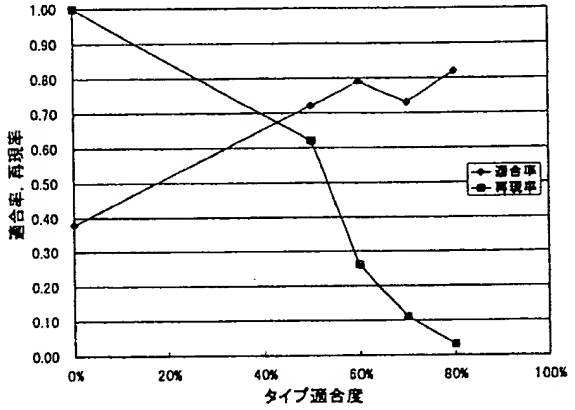


図5 『カタログ』の実験結果

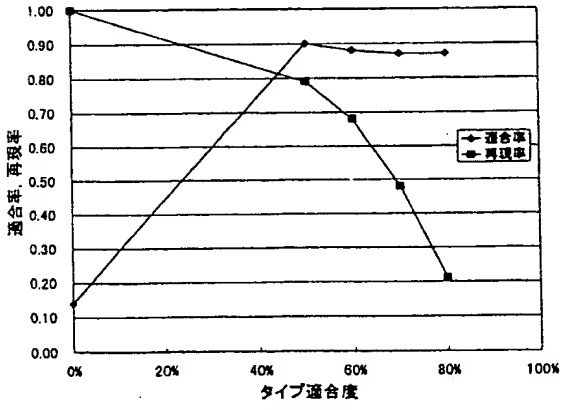


図6 『リンク集』の実験結果

図5，図6はそれぞれタイプ適合度が0以上，50以上，60以上，70以上，80以上とした場合をプロットしている．タイプ適合度が0以上というのは，文書タイプを指定しない場合と等しい．

これらの図から明らかなように，『カタログ』より『リンク集』の方が全体的に分類精度が良い．この理由は『カタログ』より『リンク集』の方が典型的なページの種類が少なく，分類もれが少なかったためであろう．『カタログ』のページには商品の製造／販売元各社の独自のスタイルがあるため，今回の実験で使用した特徴記述では拾い切れなかったページが『リンク集』より多かった．

検索件数で実験結果を示したものが図7である．

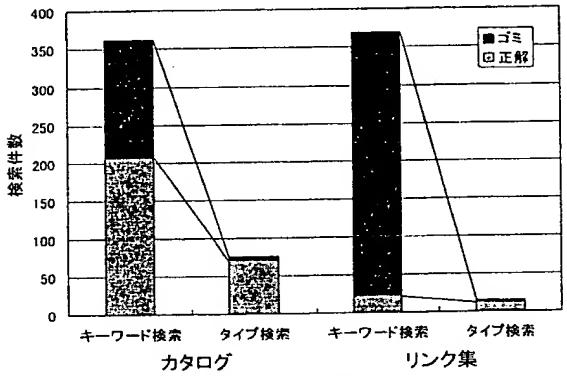


図7 検索件数

図7のキーワード検索とはタイプ適合度が0以上の場合，タイプ検索とはタイプ適合度が50以上の場合を示している．この図からも明らかなように文書タイプ検索を用いることで検索におけるゴミの混入率は1/13～1/14になっている．正解の欠落も起きているが，問題解決での利用という目的から，本システムでは，ゴミの混入率の削減の方を重視している．

4.2.2 実験2

実験1では，小規模なデータを対象に限定された文書タイプ(2タイプ)で分類精度を求めたが，本システムの有効性を検証するために更に大規模な実験を行った．

『カタログ』と『リンク集』の文書タイプについてはNETPLAZA[14]の自動収集データを用い，その他の文書タイプ(4種類)については別に収集した約15万件のデータを用いた．すべての文書タイプについて結果を表示するタイプ適合度は50以上とした．

実験では，文書タイプ毎に1単語のクエリー(基本キーワード)を用意し，その基本キーワードと文書タイプ指定，基本キーワードと文書タイプに代わる1単語キーワードの双方について検索結果の先頭20件での分類の適合率を求めた．例えば，基本キーワード"LaVie"と文書タイプ『カタログ』の論理積と，基本キーワード"LaVie"とキーワード"カタログ"の論理積である．

実験結果を図8に示す．この結果からも明らかなように，ある文書タイプのページを検索する場合，キーワードのみより本システムの方が精度が良い．各文書タイプについてそれぞれ7種類の基本キーワードで実験を行い，その平均値をプロットしているが，すべてにおいて本システムの文書タイプ指定検索の方が分類適合率が良かった．

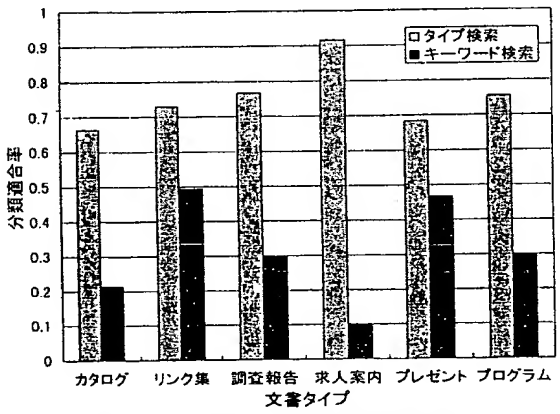


図8 6文書タイプの実験結果

実験で用いた検索システムでは，検索結果を単に文書登録の逆順で表示するようにしたため，先頭20件の比較は平均的な分類精度の比較に相当する．検索結果をタイプ適合度でソートすれば，先頭20件の精度は更に向上させることが可能である．

実験1の結果と比較すると，『カタログ』，『リンク集』ともに分類適合率が0.05～0.1程度下がっている．実験の規模の拡大に従って，各文書タイプにおけるスタイルが多様化し，1文書タイプ1特徴記述ファイルでは拾い切れないページが生じるため，適合率が下がる．しかし，その程度は非常に緩やかなものであった．他の文書タイプについてもほぼ同様の結果になることが予想できる．

このように4.1節の特徴記述作成と分類速度の結果と4.2節の分類精度の結果から，本システムは低コストで構築，維持が可能であり，実用に十分な性能を持っていることが判明した．

5. おわりに

本稿では，さまざまな問題解決に利用できる文書タイプという概念を導入したWWW検索システムについて述べた．WWWページの構造的な特徴を記述した特徴記述により，カテゴリーとは異なる視点であるWWWページの文書タイプにあらかじめ分類しておくことによって，実用的な検索性能を実現した．本システムによって今後ますます増えるであろうさまざまな検索要求に容易に対応することが可能となる．

本システムの文書タイプ検索はNETPLAZAで既に実用化されている．現在は文書タイプが2種類であるが，徐々に増やしていく予定である．

参考文献

[1] 冨田ほか：HTML文書からの商品情報抽出方式の提案，*情報処理学会第56回全国大会予稿集(3)*，pp.79-80, 1998.

[2] J. Shakes, et al: Dynamic Reference Sifting: A Case Study in the Homepage Domain, *In Proceedings of 6th WWW*, pp.189-200, 1997.

[3] R. Burke, et al: Question Answering from Frequently Ask Question Files: Experiences with the FAQ Finder System, *Univ. of Chicago, Dept. of CS*, TR-97-05, 1997.

[4] http://www.askjeeves.com/

[5] http://www.goo.ne.jp/

[6] 岩山，徳永：確率的クラスタリングを用いた文書連想検索，*自然言語処理*, Vol.5, No.1, pp.101-117, 1998.

[7] http://www.yahoo.co.jp/

[8] 村本，鷲崎：階層型知識体系を用いたWWW情報の自動カテゴリ推定方式，*情報処理学会第56回全国大会予稿集(3)*，pp.205-206, 1998.

[9] 伊藤ほか：WWWの分類・検索システムCrowww，*情報処理学会第56回全国大会予稿集(3)*，pp.221-222, 1998.

[10] H. Schutze, et al: A Comparison of Classifiers and document representation for the routing problem, *In Proceedings of 18th SIGIR*, pp.229-237, 1995.

[11] http://www.excite.com/

[12] 赤峯，福島：高速全文検索のためのフレキシブル文字列インバージョン法，*情報処理学会 Proceedings of Advanced Database Symposium '96*, pp.35-42, 1996.

[13] 松田，福島：インターネット多角的検索システムOTROS-構造的特徴量によるタイプ分類と検索-，情報処理学会第57回全国大会予稿集(3)，pp.145-146, 1998.

[14] http://netplaza.biglobe.ne.jp/keyword.html

複写される方に

情報処理学会研究報告
IPSJ SIG Notes

情処研報　Vol.99, No.20
１９９９年３月１日発行

発行所　〒108-0023 東京都港区芝浦三丁目 16 番 20 号
芝浦前川ビル　7 階

社団法人　情 報 処 理 学 会　　TEL 東京(03)5484-3535 （代表）
郵便振替口座　(00150-4-83484)

発行人　社団法人　情 報 処 理 学 会　　柳 川 隆 之
Information Processing Society of Japan